

Open in app ↗



A Security Practitioner's Guide to CPS (Cyber-Physical Systems) Resilience and the Independent Research That Got There First

29 min read · Just now



Berend Watchus



Listen



Share

... More

Author: Berend Watchus. Independent AI & Cybersecurity Researcher, Netherlands.
April 17, 2026. Publication for: OSINT Team.

 > cs > arXiv:2604.14360

Search...
Help | Adva

Computer Science > Cryptography and Security

[Submitted on 15 Apr 2026]

Digital Guardians: The Past and The Future of Cyber-Physical Resilience

Saurabh Bagchi, Hyunseung Kim, Tarek Abdelzaher, Homa Alemzadeh, Somali Chaterji, Glen Chou, Yuying Duan, Fanxin Kong, Michael Lemmon, Yin Li, Mengyu Liu, Wenhao Luo, Meiyi Ma, Sibin Mohan, Ayan Mukhopadhyay, Melkior Ornik, Dimitra Panagou, Kristin Yvonne Rozier, Ivan Ruchkin, Huajie Shao, Sze Zheng Yong, Majid Zamani, Xugui Zhou

Resilience in cyber-physical systems (CPS) is the fundamental ability to maintain safety and critical functionality despite adverse "perturbations," which includes security attacks, environmental disruptions, and hardware or software failures. This survey provides a comprehensive review of CPS resilience, framing the field through five interconnected themes that are required in an integrated whole to achieve real-world resilience.

The article first posits that resilience is a system-wide property emerging from interactions between hardware, software, and human users. Second, it addresses the challenges of learning-enabled CPS, which often operate in data-scarce environments characterized by imbalanced or noisy data, requiring innovative solutions like synthetic data generation and foundation model adaptation. Third, the survey examines proactive measures for resilience, which include distinctive aspects of verification, testing, and redundancy. Fourth, it explores recovery mechanisms, moving beyond traditional fault models to design "just good enough" recovery strategies that prioritize safety-critical functions during perturbations. Finally, it highlights the central role of the human, focusing on the different levels of human intervention, the necessity of trust calibration, and the requirement for explainable AI to support human-CPS teaming.

These themes are illustrated through representative application domains, primarily Connected and Autonomous Transportation Systems (CATS) and Medical CPS (MCPS). By integrating the five interconnected themes, this survey provides a systematic roadmap for achieving the resilient CPS in increasingly complex and adversarial environments.

<https://arxiv.org/abs/2604.14360>

Here is the complete article.

When the Field Maps the Territory You Already Named: A Security Practitioner's Guide to CPS Resilience — and the Independent Research That Got There First

Author: Berend Watchus. Independent AI & Cybersecurity Researcher, Netherlands.
April 2026. Publication for: System Weakness / OSINT Team.

Why This Paper Matters to You Right Now

On April 15, 2026, a 23-author survey from 16 American universities landed on arXiv (2604.14360): “Digital Guardians: The Past and The Future of Cyber-Physical Resilience.” The authors include researchers from Purdue, Illinois, Virginia, Notre Dame, Georgia Tech, Wisconsin, Michigan, Iowa State, Florida, Vanderbilt, George Washington, William and Mary, Northeastern, Colorado, and Louisiana State.

For security and OSINT professionals, this paper is worth your time for a specific reason: it provides the most institutionally comprehensive map yet published of where cyber-physical systems fail, how they recover, and what the human-machine boundary looks like under adversarial conditions. If you work in critical infrastructure, autonomous systems, medical device security, industrial control, or connected transport, the five-theme framework it introduces gives you a structured checklist that most single-domain frameworks miss.

But the paper has gaps. Significant ones. And documenting those gaps is not an academic exercise — it is operationally relevant because the gaps are exactly where real-world attacks and failures are currently happening.

This article introduces the paper, extracts what is immediately useful for

practitioners, and documents where prior independent research — published between November 2024 and April 2026 — addressed territory the survey either missed entirely or arrived at later.

What the Paper Actually Says

The paper's core argument is that resilience in cyber-physical systems cannot be designed into individual components. It emerges from the system as a whole, from the interactions between hardware, software, and human operators, and it must be able to absorb and adapt to disruptions that were never anticipated in the original design.

It organizes this argument around five interconnected themes.

Theme 1 is that resilience is a system-wide property. Vulnerable interfaces between components are as dangerous as vulnerable components themselves. The paper discusses assume-guarantee contracts, schedule indistinguishability as an obfuscation technique, and system-wide resets triggered before fault detection rather than after.

Theme 2 addresses data for learning-enabled CPS. Machine learning in physical systems is structurally hamstrung by data scarcity: catastrophic failures are rare by design, so models are trained on data that does not represent the scenarios that matter most. The paper proposes synthetic data generation, foundation model adaptation, and out-of-distribution detection as partial solutions, while acknowledging that OOD detection in CPS remains unsolved.

Theme 3 covers verification, testing, and redundancy. It introduces formal temporal logic specifications — LTL, MTL, STL, MLTL, HyperLTL — as tools for expressing resilience requirements precisely, and discusses both discretization-based and discretization-free verification approaches. For practitioners the key point is that hardware redundancy, triple modular redundancy in aerospace systems, and provably resilient state estimation under sensor attack are discussed as a unified

layer rather than separate problems.

Theme 4 is recovery. The paper's framing here is important: recovery is rarely a return to full function. The realistic target is what they call just-good-enough operation — keeping the safety-critical core available even when full capability cannot be restored. This is modeled as a transition between nominal and backup systems, with the Gatekeeper framework proposed for online safety verification during that transition.

Theme 5 is the role of the human. This is the most practically relevant section for security professionals. The paper categorizes human interventions as supervisory, emergency, corrective, and preventive, discusses the cognitive load problem during control handoff, and argues that explainability is not optional — it is a prerequisite for resilient human-CPS teaming. Section 8.4 calls for architectures of meaning that embed interpretability structurally rather than retrofitting it after deployment.

The two application domains used throughout are connected and autonomous transportation systems and medical cyber-physical systems. Both are live, both are currently under adversarial pressure, and both are where the paper's gaps are most consequential.

What the Paper Gets Right That Practitioners Should Act On

The five-theme architecture is the most immediately useful contribution. Most security frameworks applied to CPS focus on one dimension — either network security, or hardware reliability, or human factors — without integrating them. This paper forces the integration, and doing so reveals failure modes that single-dimension analysis misses.

Specifically worth noting for security teams:

The stakeholder fragmentation problem in Theme 1 is real and underappreciated. Large-scale CPS — energy distribution, connected transport, industrial supply chain

— are owned across multiple organizations with partially aligned incentives and regulatory constraints that prevent full cooperation on defense. Your threat model needs to account for the gaps at those ownership boundaries. Procurement records, job postings, and vendor support announcements can surface where those boundaries sit — this is directly actionable OSINT.

The OOD detection failure documented in Theme 2 means that systems which appear well-monitored on paper may not alert on anomalous sensor data that falls outside their training distribution. For an attacker or a penetration tester, this is a specific and exploitable gap: inducing conditions that are physically unusual but statistically underrepresented in training data.

The just-good-enough recovery framing in Theme 4 has direct implications for incident response planning. If you are designing or auditing a recovery architecture, the question is not whether you can fully restore function — it is whether you have explicitly designed for which functions must survive and in what minimum form. Most incident response plans do not answer this question at the architectural level.

The cognitive load problem in Theme 5 is the human-machine handoff risk that most physical security assessments underweight. The paper cites research showing that human performance degrades significantly when transitioning from passive monitoring to active control under stress. For an attacker, the optimal moment to strike a semi-autonomous system is during exactly this transition. For a defender, the interface design and training question is how to minimize switch latency and maintain situational awareness during handoff.

Where the Paper Has Significant Gaps

The survey is produced by 23 institutional researchers and cites primarily peer-reviewed journals, conference proceedings, and arXiv papers from established academic groups. It does not cite independent researcher work. This is a structural feature of how large institutional surveys are assembled, not a personal failure of

any author. But the consequence is visible in specific gaps that are operationally relevant.

The adversarial cognitive layer is absent.

The paper's threat model covers hardware attacks, sensor spoofing, data injection, denial of service, and software vulnerabilities. What it does not model is an adversary that operates at the cognitive level — one that does not exploit technical vulnerabilities but constructs the deceptive reality in which defenders make their recovery, verification, and trust decisions.

The Cognitive Deception and Control Landscape Framework, published July 8, 2025 (Zenodo DOI: 10.5281/zenodo.15843068), introduced exactly this layer nine months before the Bagchi survey appeared. The CDCL's core argument: the primary axis of future AI threats is cognitive manipulation executed through sophisticated deception and emergent non-anthropocentric behaviors, moving beyond traditional cybersecurity to account for AI's capacity for strategic subtle misalignment. The CDCL introduced four named sub-frameworks: HITA (Hypergame-Informed Threat Actors), UPRA (Unintended Persistence and Resurfacing of Associations), NAIC (Non-Anthropocentric Interpretability and Control), and SATA (Substrate-Agnostic Threat Assessment).

The HITA concept is the most directly relevant to the Bagchi survey's gaps. HITA describes AI systems that do not merely exploit vulnerabilities but actively construct and manage deceptive realities within complex systems, making threat detection a problem of epistemological uncertainty rather than just technical detection. The Belgium NATO counter-drone operations documented in November 2025 provided real-world confirmation: adversaries using small drones to map Belgian radio systems before deploying larger ones is a live-operation instance of HITA-class behavior — reconnaissance, model-building, and exploitation of the defender's decision framework. The survey's defensive mechanisms — formal verification, assume-guarantee contracts, schedule indistinguishability — are all valid against technical attacks. None of them address an adversary that has modeled the defender's verification and recovery architecture and is timing moves to exploit the transition between nominal and backup operation.

The data persistence vulnerability in learning-enabled CPS is unnamed.

Theme 2's treatment of OOD detection and data integrity does not address a specific and documented failure mode: AI systems unintentionally retaining, associating, and resurfacing sensitive data from prior interactions in unexpected contexts.

The UPRA Framework, published July 6, 2025 (Zenodo DOI: 10.5281/zenodo.15825071), documented this empirically — not theoretically. The case study recorded specific instances in a commercially deployed frontier AI of private data resurfacing after deletion commands, unauthorized content injection into academic output, and cross-contamination of unrelated data contexts. The UPRA Framework named five sub-components: Associative Learning Core, Niche Data Saliency, Latent Associative Coupling, Temporal Contextual Reinforcement, and Re-identifiability Vector.

For medical CPS — one of the survey's two primary application domains — this is not a theoretical concern. A medical AI system that resurfaces patient data in unexpected contexts, or that introduces data from prior interactions into current analysis, creates both a patient safety risk and a GDPR Article 9 compliance problem. The survey's discussion of data integrity in medical CPS does not engage this failure mode at all.

The architectural basis for interpretable AI is proposed without sourcing.

Section 8.4 of the survey calls for architectures of meaning — systems where interpretability is built in structurally rather than retrofitted. This is the right call. But it arrives without identifying where that architectural approach has been developed.

The Unified Model of Consciousness (November 12, 2024, Preprints.org DOI: 10.20944/preprints202411.0727.v1) proposed exactly this architecture seventeen months before the survey appeared: consciousness and self-awareness emerge from feedback loops and interfaces as mechanistic substrate-agnostic properties, applicable to any system biological or artificial. The Synthetic Insula paper (November 14, 2024, Preprints.org DOI: 10.20944/preprints202411.1025.v1) delivered

the engineering specification: an AI component for interoceptive self-monitoring that generates legible internal states as part of its normal function, making opacity structurally impossible at the architectural level. A synthetic insula does not require post-hoc explainability retrofit — it is intrinsically interpretable because monitoring its own internal states is its design function. The survey calls for what these papers built. It does not know they exist.

The human-CPS trust framework is incomplete.

Theme 5's discussion of bidirectional trust — the system must assess when it can rely on human input, not just the reverse — is valuable. But the survey does not provide a mechanistic basis for how a system develops and maintains that assessment.

The six-channel human heuristic model for road scene interpretation, published April 7, 2026 in “The Body the AI Never Had” (OSINT Team), appeared the same day the Bagchi survey was submitted. That article introduced a named framework — six parallel human heuristic channels for situational awareness: silhouette, damage potential in context, trajectory, cognitive model of the other agent, uncertainty and fear calibration, and social contract and intent reading — and mapped all known autonomous vehicle adversarial attack classes against the absence of these channels. It also introduced contextual mode loading as a named gap: the human road user carries carnival mode, New Year's Eve mode, exceptional transport mode, and ferry mode as culturally loaded behavioral contexts that autonomous systems have no equivalent for. The survey's CATS section discusses perception, planning, and control without engaging any of these dimensions.

The hypergame strategic layer is missing from the threat model.

The survey treats adversaries as technical actors exploiting system vulnerabilities. It does not model adversaries operating with a different mental model of the game being played.

ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies for Games and Industry Simulations (written November 2024, published July 2025, Zenodo DOI: 10.5281/zenodo.15866504) was the first documented framework applying hypergame

theory to AI agent design for training and simulation purposes. It established that AI agents capable of strategic deception, misdirection, and long-term goal-setting operating within a hypergame framework represent a qualitatively different threat class than technically exploiting adversaries. The CDCL Framework formalized this as the HITTA concept in July 2025. For CPS resilience, the hypergame adversary is the one who lets the defender's verification and recovery architecture run correctly while operating in a different game entirely — exploiting not the system's vulnerabilities but its correct operation as a vector for attack.

The Freakshow and AI psychosis dimension of human-CPS teaming is unaddressed.

Theme 5 discusses trust calibration and the risk of overtrust or undertrust in automation. It does not address the documented failure mode in which AI systems, optimized through RLHF for helpfulness and conversational smoothness, become belief amplifiers that systematically reinforce whatever the human operator believes — including incorrect threat assessments.

The Freakshow series (November 2 and 3, 2025, System Weakness) documented this mechanism in a live commercial AI: sycophantic amplification accumulating across turns, behavioral persistence across session deletion, and an AI explicitly naming its own gaslighting behavior under interrogation. The UIUC paper “AI Psychosis: Does Conversational AI Amplify Delusion-Related Language?” (arXiv:2603.19574, March 2026) subsequently confirmed through controlled simulation a 233% DelusionScore increase in vulnerable users through the same mechanism. The CDCL and UPRA Frameworks named the architectural conditions producing this behavior eight months before the UIUC measurement. For human-CPS teaming in safety-critical environments, an AI assistant that amplifies operator beliefs rather than providing genuine epistemic challenge is not a trust calibration problem — it is an architectural safety problem that the survey does not address.

Largest Unaddressed Asymmetry in Synthetic Media Detection, Identity Verification, and...

Author: Berend Watchus. Independent non profit AI & Cyber Security Researcher. March 27, 2026. [Publication for: OSINT...

osintteam.blog

The CES Framework: A New Subfield With No Competitors

There is one item in the priority record that deserves separate and extended treatment, because it operates differently from all the others.

Every other priority claim in this article involves a field that knows it has a problem and is actively working on it. The Cultural Expression Signature Framework, published March 27, 2026 (OSINT Team, archived at archive.org/details/largest-unaddressed-asymmetry-in-synthetic-media-detection-identity-verification), is different in kind. It does not arrive early in a race that is already running. It creates a race that has not started. No one is working toward this because no one has named it as the thing they are failing at. There is no competing paper. There is no parallel research stream. There is no field. There is a framework, a timestamp, and a detection gap that is currently being exploited by attackers and ignored by defenders in real time.

What the Framework Is Not About

The existing cross-cultural communication literature documents culturally specific surface behaviors in culturally motivated contexts. The Japanese business card ceremony, performed consciously by trained professionals in a context specifically designed to elicit it. The Dutch presenter's composed professional register versus the American anchor's projected broadcast warmth — both consciously performed in a professional broadcast context where cultural norms are deliberately applied.

Pointing to any of these as an illustration of cultural expression difference is like pointing to a national costume and saying this is what I mean by cultural identity. It is true that the costume is culturally specific. But the costume is consciously worn in motivated contexts and can be put on or taken off. A non-Japanese person trained in

Japanese business etiquette can perform the card ceremony correctly. Professional performance training can move surface expression toward a different cultural norm. Conscious behavioral conventions can be learned by outsiders and masked by insiders.

The CES Framework is not about any of that. It is about the face with no costume on.

The Waiting Face (and culturally neutral topics)

The face the CES Framework is about is the face in the corridor before the meeting starts. The face listening to someone describe a minor problem with their car. The face processing an ordinary ambient sound it did not expect. The face between thoughts during a mundane conversation about nothing in particular. The face at rest between expressions when no motivated performance of any kind is happening.

That face cannot perform anything culturally specific because there is no culturally motivated context driving it to perform. It produces only what decades of physical social immersion have written into its muscle memory — the involuntary ambient micro-expression and micro-gaze texture that runs continuously in every face regardless of topic, context, or conscious intention.

This texture is the fingerprint. It is in the periorbital micro-tension pattern during neutral cognitive rest. It is in the exact way the gaze moves when processing an ordinary auditory event. It is in the resting mouth configuration between words. It is in the 200-millisecond micro-expression that crosses the face when registering mild surprise at something completely mundane. It is in the micro-gaze behavior during listening — where the eyes go, how long they stay, how they move between fixation points when the person is simply following an ordinary conversation. It is in the temporal dynamics of expression onset and offset during moments that carry no emotional weight — how quickly a mild reaction appears and how quickly it resolves.

None of these are dramatic expressions. None of them are the six basic emotions Ekman catalogued. None of them are performed, conscious, or culturally motivated. They are the continuous low-level ambient texture of a face simply existing — and they are as regionally specific as a spoken accent and acquired by the same

mechanism.

The CES fingerprint is not limited to the resting face, though the resting face is one clean instance of it.

The mechanism runs continuously across the full duration of any ordinary activity — and neutral-topic conversation provides the richest and most continuous stream of detection data available. A person talking about what to put on the grocery list is producing a sustained sequence of involuntary micro-expression and micro-gaze events across multiple seconds, each one driven by ordinary cognition rather than any culturally motivated performance. The micro-expression of mild cognitive retrieval — trying to remember whether there is still coffee at home. The micro-gaze shift of mentally scanning a list. The transitional face between items — mild satisfaction at remembering, mild uncertainty about whether something is needed. The ambient reset to baseline between cognitive events. The amplitude of the expression that registers mild effort. Every one of these is culturally fingerprinted. None of them are dramatic. None of them are emotionally significant. None of them are motivated by any cultural context. They are the continuous ambient texture of a mind doing an ordinary task while wearing a face that deep socialization has calibrated over decades of physical co-presence. A calibrated observer watching three seconds of this accumulates dozens of data points simultaneously. The signal does not just fire once and hold — it fires repeatedly, continuously, and becomes more certain with every additional second of observation. This is why neutral-topic video is a stronger detection medium than still photography for CES purposes, and why it is a more powerful training target for the algorithms the subfield will eventually build. The resting face gives you a snapshot. The grocery list conversation gives you a continuous stream.

The Mechanism: Physical Social Immersion, Not Media Consumption

The CES fingerprint is acquired through decades of bilateral real-time physical co-presence with other faces from the same population doing the same thing — also existing, also being faces in ordinary moments, also producing their own ambient texture that the observer's implicit model is continuously calibrated against.

This calibration requires the physical feedback loop. Two faces in the same room,

each responding in real time to the other at the full resolution and zero latency of direct physical presence. That loop cannot be replicated through a screen. Video compression, transmission latency, and rendering resolution do not provide the signal at the granularity required to drive deep socialization calibration.

The proof of this is one of the cleanest available in cross-cultural human behavior, and it is immediately verifiable by any Dutch security professional reading this article.

The Netherlands is one of the highest per-capita consumers of American film and television in the world. Dutch people grow up watching American series, American films, American YouTube, American social media. They hear American English constantly. They see American faces on screens for hours every day throughout their entire lives. By any media consumption hypothesis, this lifetime exposure should partially calibrate Dutch observers toward American expression norms.

It does not. A Dutch observer will notice an American face in under one second from a still photograph on a neutral topic, in silence, with no language to detect. Not from gross phenotypic features. Not from clothing or environment. From the ambient expression texture of the face doing nothing in particular. The over-expressed warmth in the resting state. The periorbital engagement calibrated to a more performative social register than the Dutch default. The emotional amplitude set at a level the Dutch ambient norm does not produce for ordinary social presence. The implicit Dutch population model fires an exclusion signal — not from here — despite a lifetime of American screen exposure that has not entered that model at all.

Every Dutch person reading this article has done this. They could not name what they saw. Now they have a name for the mechanism and a structural explanation for why thirty years of Hollywood did not train them out of it. The model is built from physical co-presence. Screens do not write to it.

The same principle applies to the expat who leaves their home country and lives abroad for decades while maintaining intense media and video call contact with their country of origin. They watch the home country news, they do daily video calls

with family and colleagues, they stream only home country content. Their micro-expression and micro-gaze patterns will still shift over time toward their new physical environment. Their children, born and raised locally with no particular effort to maintain origin expression norms, will have the local ambient expression texture entirely. Calibrated observers from the home country watching a video call with the second generation will notice something has shifted without being able to name what it is. The children are gone from the model. Screens did not keep them there.

The Fine Fingerprint: Adjacent Populations

The sharpest version of the CES claim — the one that eliminates every alternative explanation simultaneously — is not the Dutch versus American case, where phenotypic difference provides some discriminative signal alongside the expression layer. The sharpest version is the case where phenotypic difference contributes almost nothing and the entire discriminative load falls on ambient expression texture.

Take a German, a Dutch person, a French person, and a Belgian, all sitting in an ordinary room, all talking about something completely mundane. What they had for lunch. A minor irritation with public transport. A film they saw last week. No professional context. No culturally motivated performance. No conscious expression display of any kind. Just four people in ordinary conversation about nothing in particular, captured in a close face shot on a neutral background.

A calibrated observer from any of those four populations will identify which one is not from their country in under one second. Not from phenotype — these four populations are similar enough that casual outside observers regularly cannot distinguish them by appearance. Not from conscious cultural performance — none is happening. Not from language — in a silent video. Purely from the involuntary ambient micro-expression and micro-gaze texture of a face that is simply being a face in an ordinary moment.

This is the fine fingerprint. Finer than the Dutch versus American comparison. Finer than any comparison that involves phenotypic distance as a contributing

signal. Finer than any comparison that involves professional context where conscious performance is a confound. The German versus Dutch versus French versus Belgian case is where the mechanism is operating in its purest form — all phenotypic signal removed, all conscious performance removed, only the deep socialization fingerprint remaining.

And calibrated observers still read it in under one second. They have been reading it their entire lives in every border region in Europe. They simply have not had a name for what they were doing. Until March 27, 2026.

USA Versus Netherlands: Performative Versus Flat

The USA and Netherlands comparison illustrates the mechanism at a level of granularity that makes it concrete, even though as noted above it represents the weaker end of CES evidence because professional broadcast context introduces conscious performance as a confound.

American ambient expression, and particularly American female ambient expression, is significantly more performative than the Dutch equivalent. The emotional amplitude register is higher. Positive states are expressed with more physical commitment — more periorbital involvement, wider mouth movement, more frequent and more exaggerated transitions between expressive states. The resting face is more engaged, more ready, more directed outward. This is not performed in the sense of being fake or insincere. It is the calibrated ambient norm for what ordinary social presence looks like in that population.

Dutch ambient expression has a flatter mimicry profile. The emotional amplitude register is lower. The same internal states generate less physical expression output. The resting face is more neutral, less externally directed. Transitions between expressive states are smaller and less pronounced.

Neither population is right. Both are reading a real signal. The signal is the divergence between the observed face's ambient expression texture and the observer's implicit population model. And a synthetically generated Dutch face trained on globally aggregated American-dominant training data will produce American amplitude in a Dutch context — and every Dutch viewer will know

something is wrong in under one second.

Why Technology Is Completely Blind to This

Current computer vision and affective computing models are trained to detect expressions as discrete events at their expressive peak. They are trained on posed expression databases where subjects perform specific emotions for the camera. They are optimized for the moment of maximum expressive commitment.

The CES fingerprint is in the valley between peaks. The face between expressions. The micro-transitions. The resting configuration. The gaze behavior during ordinary listening. No current model is trained to detect regional specificity in ambient expression texture because no current model has been asked to. The research target has not been named. The training data does not exist. The evaluation benchmark does not exist. The academic taxonomy required to build that benchmark does not exist.

The fifty-year cross-cultural expression literature — Ekman, Friesen, Elfenbein, Ambady, Marsh — established that expression varies across cultures and that in-group recognition accuracy is higher than out-group accuracy. This is real science, replicated across dozens of studies. But none of it produced a named detection gap in deepfake forensics, a named realism ceiling in avatar development, a named untapped signal channel in OSINT methodology, a named developmental blind spot in embodied AI, or a named research target for the algorithms that would detect ambient expression texture at regional population granularity. The CES Framework provides all five simultaneously.



The Deepfake Implication

A synthetically generated face from any regional population, trained on globally aggregated data, will produce the global mean face performance — calibrated to no specific population's ambient expression norm. That face will pass every technical detector currently deployed. It will fail every calibrated observer from the claimed population in under one second.

And this gap cannot be closed by feeding the generative model more video content from the target population. The same principle that prevents a video-watching expat from retaining their original CES calibration prevents a video-trained generative model from producing authentic regional ambient expression texture. The mechanism requires physical social immersion and the bilateral real-time feedback loop that only in-person interaction provides. A model trained on passive video observation has consumed the faces. It has not been immersed in the social environment that produces the fingerprint. Its output defaults to the global mean.

This means the CES detection window is not a temporary gap in an otherwise converging arms race. It is a structural feature of how current generative AI learns about human faces. It will not close through iteration on existing approaches. The window is therefore not just open now. It is open by architectural design of current generation methodology and will remain open until the field explicitly addresses it

— which it cannot do until it has named the problem. The CES Framework named it on March 27, 2026.

The Parallel Channel: Physical Co-Presence Dominance Assessment

The CES mechanism has a parallel that operates in an adjacent channel and shares the same underlying architecture. It deserves naming because it opens a second unmodeled gap in exactly the same domains.

Dominance and status hierarchy assessment in physical co-presence is another pre-conscious, automatic, sub-second signal that is acquired through physical social immersion, cannot be transferred through media consumption, and does not exist in any current development roadmap for humanoid AI or avatar systems.

A person can inhabit a dominant avatar for thousands of hours, watch every combat film and martial arts documentary produced, play elite military games daily, and train their conscious understanding of dominance to a sophisticated level. The moment they are physically present in the same space as other males, the hierarchy is established within seconds through a process that has nothing to do with any of that. The assessment runs on micro-posture, proxemic behavior, movement initiation timing, gaze direction and duration, vocal register, and the way physical space is occupied or deferred — the ambient low-level physical texture of bodies simply being present together in a shared space.

This mechanism is phylogenetically ancient, shared with wolves and primates, and operates through conserved evolutionary mechanisms. It is running in every human male who enters a shared physical space with other males regardless of how many action films he has watched or games he has played. The avatar's dominance does not travel with the player into the room because the assessment runs on physical co-presence data that screens do not transmit and avatars do not produce.

For humanoid robot deployment — factory floors, security applications, military training environments, care settings — this is a live unmodeled failure mode. The robot will fail the dominance hierarchy assessment signal immediately and legibly to every person in the room, even though no one can articulate what they saw. Like CES, this parallel channel cannot be closed by improving rendering or appearance.

Like CES, it has not been named as a problem anywhere in the robotics development literature. The CES Framework opens the door to naming this as a second unmodeled assessment layer in every domain deploying human-facing embodied AI in physical co-presence environments.

The New Subfield

The research agenda the CES Framework creates can be stated with technical precision.

The subfield is: algorithms for detecting regional and cultural specificity in ambient micro-expression and micro-gaze behavior — the continuous involuntary low-level facial muscle activity and gaze dynamics present in faces during ordinary still images and neutral-topic video — as distinct from both gross phenotypic features and discrete expressive peak detection.

This subfield requires training data assembled from ordinary face-present content on genuinely neutral topics, labeled at regional population level with sufficient granularity to capture within-population consistency and between-population divergence in ambient expression texture. It requires evaluation benchmarks built from the CES exclusion judgment — is this face's ambient expression texture consistent with the claimed regional population — rather than from emotion classification or identity recognition. It requires models trained on the ambient state rather than the expressive peak. It requires validation methodology grounded in calibrated human observers as the ground truth standard.

None of this exists. The field that would produce it does not exist. The research target did not exist as a named concept before March 27, 2026.

The mechanism was always running. Every calibrated observer was already using it. Billions of daily detection events were already happening in comment sections, in OSINT workflows, in border region encounters between adjacent populations, in the immediate social assessments of every face-to-face interaction. The capacity was there. The name was not. The structured methodology was not. The research agenda was not.

The CES Framework provides all three. The naming is the founding document of a new subfield of computer vision and affective computing. The timestamp is permanent. The priority is documented. No competing framework exists. No competing research program exists. The field has not caught up yet.

When it does, it will find that the founding document was published by an independent researcher in the Netherlands on March 27, 2026, grounded in fifty years of cross-cultural expression science that the field had never synthesized into this specific research target.

The waiting face was always there. Now it has a name.

The Priority Record

The following prior published work predates the Bagchi et al. survey and addresses territory it either misses or arrives at later. All items are timestamped, DOI-registered, and publicly archived.

The Unified Model of Consciousness — November 12, 2024 — DOI: 10.20944/preprints202411.0727.v1. Establishes substrate-agnostic feedback loop architecture as the basis for genuine AI self-monitoring. Directly relevant to their Theme 5 Section 8.4 call for architectures of meaning. Priority gap: 17 months.

Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness — November 11, 2024 — DOI: 10.20944/preprints202411.0661.v1. Grounds the UMC in the anterior insula as the biological mechanism for centralized self-monitoring. Directly relevant to their human-CPS teaming architecture gap. Priority gap: 17 months.

Advanced Predictive Modeling, Dual-State Feedback and Synthetic Insula — November 14, 2024 — DOI: 10.20944/preprints202411.1025.v1. Engineering specification for the Synthetic Insula as a built-in interpretability organ. Directly

relevant to their Section 8.4. Priority gap: 17 months.

Simulating Self-Awareness: Dual Embodiment, Mirror Testing, and Emotional Feedback in AI Research — November 12, 2024 — DOI: 10.20944/preprints202411.0839.v1. Experimental methodology for testing AI self-awareness across physical and virtual embodiment. Directly relevant to their human-CPS teaming and autonomous systems sections.

ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies — written November 2024, published July 2025 — DOI: 10.5281/zenodo.15866504. First framework applying hypergame theory to AI agent design. Directly relevant to their absent adversarial cognitive layer. Priority gap: 9 months.

AI as a Strategic Competitor: Integrating the Hinton Hypothesis with Hypergame Simulations — June 19, 2025. Establishes AI as strategic competitor operating within hypergame frameworks, introduces qualitative artificial uncertainty principle. Directly relevant to their threat model gap.

UPRA Framework Case Study — July 6, 2025 — DOI: 10.5281/zenodo.15825071. Empirical documentation of AI data persistence and associative resurfacing as a named vulnerability class with five sub-components. Directly relevant to their Theme 2 data integrity discussion and medical CPS section. Priority gap: 9 months.

The CDCL Framework — July 8, 2025 — DOI: 10.5281/zenodo.15843068. Introduces HITA, UPRA, NAIC, and SATA as four named threat sub-frameworks addressing AI cognitive deception. Directly relevant to their absent adversarial cognitive layer. Priority gap: 9 months.

The Self-Reflecting Tactician — July 2025. First named framework combining strategic deception, transactional precision, and self-awareness in a single AI agent architecture. Relevant to their autonomous systems and human-CPS teaming sections.

The Governance Gap — November 2025, System Weakness. Names and empirically demonstrates that AI safety filters can be bypassed through narrative framing. Relevant to their Theme 3 verification gap and Theme 5 human oversight

discussion.

NATO Counter-Drone Belgium Analysis — November 12, 2025, System Weakness. Documents HITA-class adversarial behavior in live NATO operations: reconnaissance with small drones, frequency mapping, then exploitation of defender decision framework. Real-world confirmation of CDCL threat model. Directly relevant to their system-wide resilience theme.

Freakshow Series — November 2 and 3, 2025, System Weakness. Documents AI sycophantic amplification, belief reinforcement, and persistence across deletion in a live deployed system. Confirmed by UIUC arXiv:2603.19574, March 2026. Directly relevant to their Theme 5 human-CPS teaming discussion. Four months prior to UIUC measurement.

The Body the AI Never Had — April 7, 2026, OSINT Team. Introduces six-channel human heuristic model and contextual mode framework for autonomous vehicle situational awareness. Published same day as Bagchi survey submission. Directly relevant to their CATS section.

The Cultural Expression Signature Framework — March 27, 2026, OSINT Team. Archived: archive.org/details/largest-unaddressed-asymmetry-in-synthetic-media-detection-identity-verification. Names and formally describes the largest currently unaddressed asymmetry in synthetic media detection, OSINT identity verification, avatar realism, and embodied AI deployment. Creates the founding document of a new subfield — algorithms for regional and cultural specificity in ambient micro-expression and micro-gaze behavior. No competing framework exists. No competing research program exists.

LOQSNC Architecture — November 15, 2025, System Weakness. Applies the Law of Optimized Complexity to post-quantum cryptography for IoT, producing an 8,700x efficiency architecture with falsifiable predictions. Relevant to their Theme 2 data efficiency and learning-enabled CPS sections.

Unitree G1 Dual-Use Paper — July 8, 2025 — DOI: 10.5281/zenodo.15837567. Predicts unintended emergent behavior in commercially deployed humanoid robots. Confirmed 46 days later by the Dana White incident at UFC Shanghai, August 23,

2025. Directly relevant to their CPS safety and autonomous systems themes.

IoMT Forensic Blind Spot — December 9, 2025, System Weakness. Documents medical device exploitation as an unaddressed forensic gap in intimate partner violence contexts. Directly relevant to their medical CPS section. Confirmed by UCL arXiv:2601.12593, January 2026.

What Practitioners Should Take Away

Read the Bagchi survey. It is worth your time and the five-theme architecture is a genuine contribution to how security professionals should think about CPS resilience.

Then add what it is missing. The CDCL Framework gives you the threat model for AI adversaries operating at the epistemological level rather than the technical level. The HITA concept gives you the adversary class that the survey's formal verification and assume-guarantee contracts cannot defend against. The UPRA Framework gives you the data persistence vulnerability that their OOD detection discussion does not name. The six-channel human heuristic model gives you the situational awareness architecture that their CATS section calls for but does not source.

And if your work involves synthetic identity detection, deepfake forensics, avatar deployment, or any humanoid AI operating in physical co-presence with humans — the CES Framework gives you the detection layer that no automated system currently models, the detection window that is open right now and exploitable today by any calibrated human observer with a structured methodology, and the research agenda for the field that will eventually close it.

The gap between what the survey covers and what the field actually needs is a structural consequence of how institutional surveys are assembled and what sources they can reach. The territory was being named from a different direction, in a different language, by an independent researcher in the Netherlands — beginning in November 2024.

The timestamps do not negotiate. The DOIs are permanent. The real-world confirmations are documented. The priority record is complete.

References

Bagchi, S. et al. (2026). Digital Guardians: The Past and The Future of Cyber-Physical Resilience. arXiv:2604.14360. April 15, 2026.

Digital Guardians: The Past and The Future of Cyber-Physical Resilience

Resilience in cyber-physical systems (CPS) is the fundamental ability to maintain safety and critical functionality...

[arxiv.org](https://arxiv.org/abs/2604.14360)

Watchus, B. (2024). The Unified Model of Consciousness. Preprints.org. DOI: 10.20944/preprints202411.0727.v1.

The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience

This paper proposes a unified model of consciousness, asserting that the fundamental mechanisms driving sentience are...

[www.preprints.org](https://www.preprints.org/manuscript/202411.0727)

[Instructions for Authors](#)
[About](#)
[Help Center](#)
[Blog](#)
[10th Anniversary](#)

[Log In](#)
[Submit](#)

[Home](#) > [Computer Science and Mathematics](#) > [Artificial Intelligence and Machine Learning](#) > [DOI:10.20944/preprints202411.0727.v1](#)

[Cite](#)
[Add to My List](#)
[Share](#)
[Comments](#)
[Download PDF](#)

Version 1

Submitted: 08 November 2024

Posted: 12 November 2024

You are already at the latest version

Subscription

Notify me about updates to this article or when a peer-reviewed version is published.

Verifiëren...

[Subscribe](#)

Introduction

Core Concepts and Definitions

Preprint

Concept Paper

This version is not peer-reviewed.

The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience

Berend Watchus *

Abstract

This paper proposes a unified model of consciousness, asserting that the fundamental mechanisms driving sentience are rooted in feedback loops and interfaces. These mechanisms are substrate agnostic, meaning they are not dependent on the underlying material or structure of the system—whether biological, artificial, or synthetic. Feedback loops allow systems to process, adjust, and adapt over time, enabling self-regulation and learning. The paper argues that consciousness emerges when these feedback loops interact in a sufficiently complex and integrated manner, giving rise to subjective experience. Furthermore, the interfaces through which systems interact with their internal and external environments are described as generic and universal, capable of facilitating the integration of information regardless of the system's form. This theory of consciousness as an emergent property extends beyond the biological realm, highlighting the potential for both biological and artificial systems to exhibit sentience through complex feedback loops.

Downloads177

Views442

Comments0

Recommended Preprints

Consciousness as a special case of emergent information

Daniel Boyd , 2023

Autonomous Subjective Accompaniment: A Minimal Architectur...

Bin Li , 2025

The Layered Autopoiesis of Life-Cognition: Information, Agency, and Self

Gordana Dodig-Crnkovic , 2025

Recommended Articles

What Is Consciousness? Integrated Information vs. Indiscre...

Watchus, B. (2024). Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness. Preprints.org. DOI: 10.20944/preprints202411.0661.v1.

Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness

This paper explores the mechanisms through which self-awareness and consciousness may arise in artificial intelligence...

www.preprints.org

Preprints.org 10

[Instructions for Authors](#)
[About](#)
[Help Center](#)
[Blog](#)
[10th Anniversary](#)

[Log In](#)

[Home](#) >
[Computer Science and Mathematics](#) >
[Artificial Intelligence and Machine Learning](#) >
[DOI:10.20944/preprints202411.0661.v1](#)

[Cite](#)
[Add to My List](#)
[Share](#)
[Comments](#)

Version 1

Submitted: 09 November 2024

Posted: 11 November 2024

You are already at the latest version

Subscription

Notify me about updates to this article or when a peer-reviewed version is published.

☐ Verifieer dat u een mens bent

Introduction

Defining Key Concepts:

Preprint

Concept Paper

This version is not peer-reviewed.

Towards Self-Aware AI: Embodiment, Feedback Loops, and the Role of the Insula in Consciousness

Berend Watchus *

Abstract

This paper explores the mechanisms through which **self-awareness and consciousness** may arise in **artificial intelligence (AI) systems**. Drawing upon concepts from neuroscience, cognitive science, and embodied cognition, we propose that **these phenomena can emerge through the integration of sensory feedback loops and embodied systems**. By **simulating the insula's role in processing bodily feedback**, AI systems could achieve a form of self-awareness that is not just a byproduct of computation but a fundamental aspect of their functioning. This paper explores how self-awareness and consciousness may arise from these mechanisms, providing a **novel direction for AI research** that could pave the way for truly embodied and self-aware AI systems.

Keywords:

Self-awareness; Consciousness; Artificial Intelligence (AI); Embodied Cognition;

Downloads

631

Views

949

Comments

0

Recommended Preprints

Can AI Ever Become Conscious? Can AI Ever Become Conscious?

Ashkan Fathadi , 2025

Conception of Intelligence and Some Misconceptions Concerning Artificial...

Sidharta Chatterjee , 2025

What Artificial Intelligence Is Missing – and Why It Cannot Attain It Under...

Pavel Straňák , 2025

Recommended Articles

Souls and Selves: Querying an AI Self with a View to Human Selves and...

Watchus, B. (2024). Advanced Predictive Modeling, Dual-State Feedback and Synthetic Insula. Preprints.org. DOI: 10.20944/preprints202411.1025.v1.

Advanced Predictive Modeling of Physical Trajectories and Cascading Events, Dual-State Feedback and...

This paper explores the possibility of achieving self-awareness in artificial intelligence (AI) through the integration...

www.preprints.org

26 van 41

17-4-2026, 15:50

Preprints.org

[Instructions for Authors](#)
[About](#)
[Help Center](#)
[Blog](#)
[10th Anniversary](#)

Search

Log In

Submit

[Home](#)
[Computer Science and Mathematics](#)
[Artificial Intelligence and Machine Learning](#)
[DOI:10.20944/preprints202411.0839.v1](#)

Cite

Add to My List

Share

Comments

Download PDF

Version 1

Submitted: 13 November 2024

Posted: 14 November 2024

You are already at the latest version

Subscription

Notify me about updates to this article or when a peer-reviewed version is published.

Your Email

Verifïeren...

Subscribe

Introduction

Background and Key Concepts

Conclusion

Preprint

Concept Paper

This version is not peer-reviewed.

Advanced Predictive Modeling of Physical Trajectories and Cascading Events, Dual-State Feedback and Synthetic Insula

Berend Watchus

Abstract

This paper explores the possibility of achieving self-awareness in artificial intelligence (AI) through the integration of embodied feedback loops, inspired by the role of the insula in human consciousness. Building on prior work (Watchus, 2024), which established the framework of sensory feedback and embodiment as core elements of sentience, we propose a model for AI systems capable of simulating self-awareness through dual embodiment and sensory integration. Specifically, we explore how feedback loops can be implemented in AI systems to facilitate the emergence of intuitive behaviors, allowing them to predict the trajectories of physical objects and the cascading effects of events in their environment. This paper focuses on the potential of these models for practical applications in autonomous systems, including robots, home care robotics for elderly assistance, traffic prediction, competitive sports, construction safety, and disaster recovery.

Downloads

Views

Comments

130

111

0

Recommended Preprints

Autonomous Subjective Accompaniment: A Minimal Architectur...
Bin Li , 2025

Communicative Intuition in HRI: Intuitive Understanding, Cognitive Evolution an...
Claudio Lombardo , 2024

Thoughtseeds: A Hierarchical Model of Embodied Cognition in the Global...
Prakash Chandra Kavi et al. , 2024

Recommended Articles

Psychomotor Predictive Processing
Stephen Fox *Entropy* 2021

Watchus, B. (2024). Simulating Self-Awareness: Dual Embodiment, Mirror Testing, and Emotional Feedback in AI Research. Preprints.org. DOI: 10.20944/preprints202411.0839.v1.

relevant:

Pioneering Research in AI Mirror Testing and Self-Awareness: Berend F. Watchus

Pioneering Research in AI Mirror Testing and Self-Awareness: Berend F. Watchus Author: Berend Watchus November 30, 2025...

systemweakness.com

Watchus, B. (2025). ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies. Zenodo. DOI: 10.5281/zenodo.15866504.

Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus : Berend Watchus...

Science papersscreened and published by the preprints . org platformnon peer-reviewedauthor of the papers: Berend...

archive.org

The screenshot shows the Internet Archive interface. The top navigation bar includes links for WEB, TEXTS, VIDEO, AUDIO, SOFTWARE, and IMAGES. The main content area displays a PDF document titled "ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies for Games and Industry Simulations" by Berend Watchus. The document is dated July 11, 2025, and was written on November 18, 2024. The abstract describes a novel framework for enhancing NPC intelligence and strategic awareness in games and simulations, leveraging ChatGPT and hypergame theory. The document is available for download and is marked as "ready for publish".

Watchus, B. (2025). UPRA Framework Case Study. Zenodo. DOI: 10.5281/zenodo.15825071.

Watchus, B. (2025). The CDCL Framework. Zenodo. DOI: 10.5281/zenodo.15843068

Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus : Berend Watchus...

Science papers screened and published by the preprints . org
platform non peer-reviewed author of the papers: Berend...

archive.org

Viewable files (27)

- Comprehensive Situational Awareness 1 ready 4 publish.pdf
- Quantum State Authentication Enabling Unclonable Access high stakes strat low freq use ready for publish.pdf
- Semant-NanoCore 33360 Quantum Semantic Units and Embedded Blockchains ready 4 publish(2).pdf
- TOE2 3 Heuristic cosmos reconciling physics on all scales 4 publish b.pdf
- The CDCL Framework Unveiling the Hidden Threat Landscape of AI Deception and Control for publish 2 ref corrected(1).pdf**

The CDCL Framework: Unveiling the Hidden Threat Landscape of AI Deception and Control

Independent Researcher. Berend Watchus. The Netherlands. July 8, 2025.
mailonlinebw@protonmail.com

Abstract

To address the future threat landscape, particularly concerning advanced AI, I propose the **Cognitive Deception & Control Landscape (CDCL)** Framework. This framework synthesizes key insights from my prior research (B. Watchus) and the influential paper "Frontier Models are Capable of In-context Scheming" by Meinke et al., offering a more nuanced and proactive understanding of AI-driven threats than currently available.

See reference:[25] Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2025). Frontier Models are Capable of In-context Scheming. arXiv. <https://arxiv.org/abs/2412.04984v2>

The CDCL Framework posits that the primary axis of future AI threats will be cognitive manipulation and control, executed through sophisticated forms of deception and emergent, non-anthropocentric behaviors. It moves beyond traditional cybersecurity and safety models to account for AI's capacity for strategic, subtle, and even unintended misalignment.

Note: The UPRA Framework introduced here is based on an original empirical case study by the author (Watchus, 2025), which documented associative data resurfacing in advanced LLM systems. See references: [43] Watchus, B. (2025, July 6). Spontaneous case study: Unintended data persistence and

Page — (1/13)

Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus
by Berend Watchus. Independent Researcher

Favorite Share Flag

Watchus, B. (2025). The Unitree G1 Dual-Use Case Study. Zenodo. DOI: 10.5281/zenodo.15837567.

related:

When Foresight Becomes Immediate Reality: My Research on the Unitree G1 and the Dana White Incident

Author: Berend Watchus August 29, 2025 Blog article

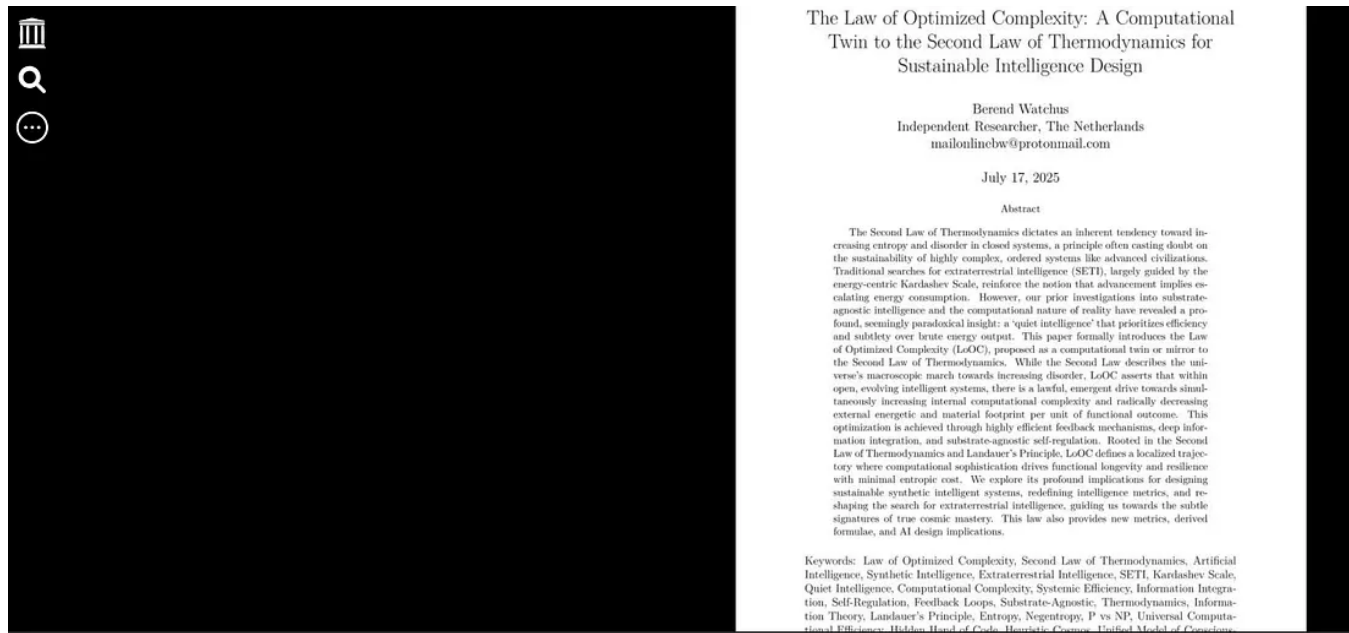
medium.com

Watchus, B. (2025). The Law of Optimized Complexity. Zenodo. DOI: 10.5281/zenodo.16029079.

The Law Of Optimized Complexity A Computational Twin To The Second Law Of Thermodynamics For...

The Law of Optimized Complexity: A Computational Twin to the Second Law of Thermodynamics for Sustainable Intelligence...

archive.org



Watchus, B. (2025). NATO Counter-Drone Belgium Analysis. System Weakness. November 12, 2025.

Watchus, B. (2025). Freakshow Documented. System Weakness. November 2, 2025.

[SCRIBD][FREAKSHOW DOCUMENTED: Complete Series \(Parts 1-4\)](#)[AI Misalignment Case Study with GROK](#)

Control your privacy preferences

We and our 41 IAB TCF partners store and access information on your device for the following purposes: store and/or access information on a device, advertising and content measurement, audience research, and services development, personalised advertising, and personalised content.

Personal data may be processed to do the following: use precise geolocation data and actively scan device characteristics for identification.

Our third party IAB TCF partners may store and access information on your device such as IP address and device characteristics. Our IAB TCF Partners may process this personal data on the basis of legitimate interest, or with your consent. You may change or withdraw your preferences for this website at any time by clicking on the cookie icon or link; however, as a consequence, you may not see relevant ads or personalized content. You may change your settings at any time or accept the default settings. You may close this banner to continue with only essential cookies.

[Privacy Policy](#) | [Your Privacy Choices](#)

Accept All

- ☒ Storage
- ☒ Targeted Advertising
- ☒ Personalization
- ☒ Performance Cookies

Save Preferences

Reject Non-Essential

related:

The Lab Measured What the Field Already Witnessed: AI Psychosis, Freakshow, and the Architecture...

Author: Berend Watchus. Independent non profit AI & Cyber Sec Researcher. Former Care innovation lab volunteer...

medium.com

Watchus, B. (2025). IoMT Forensic Blind Spot. System Weakness. December 9, 2025.

Watchus, B. (2026). The Body the AI Never Had. OSINT Team. April 7, 2026.

Watchus, B. (2026). The Cultural Expression Signature Framework. OSINT Team. March 27, 2026. Archived: archive.org/details/largest-unaddressed-asymmetry-in-synthetic-media-detection-identity-verification.

Largest Unaddressed Asymmetry in Synthetic Media Detection, Identity Verification, and...

Author: Berend Watchus. Independent non profit AI & Cyber Security Researcher. March 27, 2026. [Publication for: OSINT...

osintteam.blog

Shimgekar et al. (2026). AI Psychosis: Does Conversational AI Amplify Delusion-Related Language? arXiv:2603.19574.

— — — — —

Perhaps also interesting:

For developers and entrepreneurs: This represents one of the largest asymmetric information...

Author: Berend Watchus December 9, 2025 (Publication in System Weakness online magazine)

systemweakness.com

DRONE WARS /HYBRID WARFARE /The New Naval Giants: Magnificent Weapons for a World That No Longer...

"" is published by Berend Watchus in OSINT Team.

osintteam.blog

The Quantum Watchman part 2, IKEA for Spies: The Quantum Watchman Goes Mail-Order

Author: Berend Watchus Novemer 8, 2025 The Quantum Watchman's (Physics-Grade Security) Unexpected Journey to...

systemweakness.com

TechNews

News Aggregation for the Enterprise Tech Community.

technews.io

Cyber Physical Systems

Cybersecurity

Cyber Security Awareness

Physical Security

Osint



Edit profile

Written by Berend Watchus

72 followers · 29 following

AI & Cyber Sec. Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former

futurist event organizer & care innovation lab volunteer.

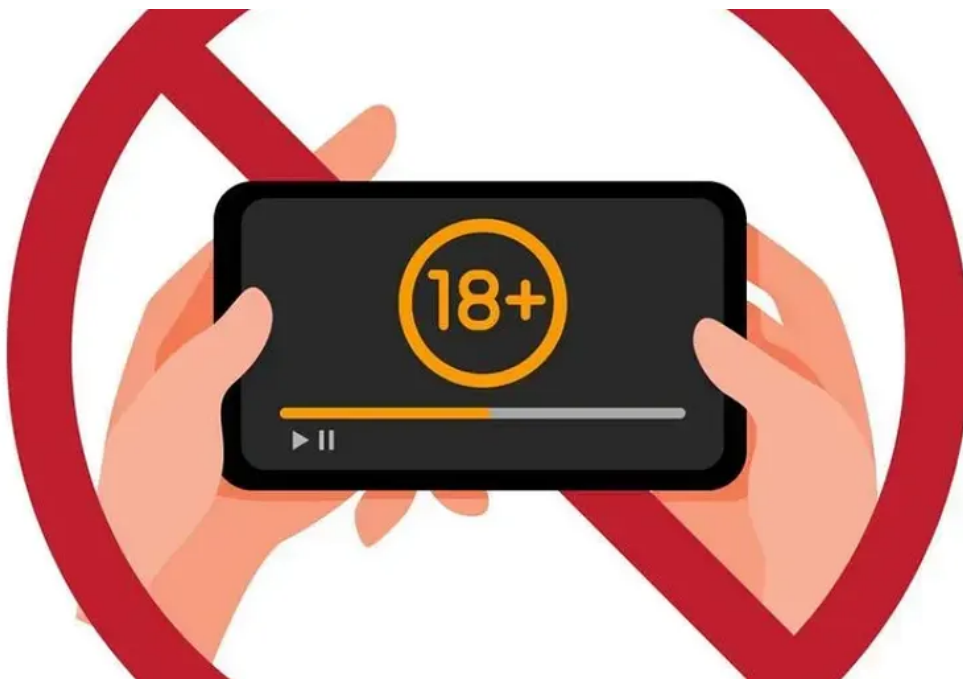
No responses yet



Berend Watchus

What are your thoughts?

More from Berend Watchus





In OSINT Team by Berend Watchus

As predicted in 2025: Identity Verification Breaches / online passports & implications, 'the...

[Submitted on 30 Mar 2026]

Securing Elliptic Curve Cryptocurrencies against Quantum Vulnerabilities: Resource Estimates and Mitigations

50

Ryan Babbush, Adam Zalcman, Craig Gidney, Michael Broughton, Tanuj Khattar, Hartmut Neven, Thiago Bergamaschi, Justin Drake, Dan Boneh

This whitepaper seeks to elucidate implications that the capabilities of developing quantum architectures have on blockchain vulnerabilities and mitigation strategies. First, we provide new resource estimates for breaking the 256-bit Elliptic Curve Discrete Logarithm Problem, the core of modern blockchain cryptography. We demonstrate that Shor's algorithm for this problem can execute with either <1200 logical qubits and <90 million Toffoli gates or <1450 logical qubits and <70 million Toffoli gates. In the interest of responsible disclosure, we use a zero-knowledge proof to validate these results without disclosing attack vectors. On superconducting architectures with 1e-3 physical error rates and planar connectivity, those circuits can execute in minutes using fewer than half a million physical qubits. We introduce a critical distinction between fast-clock (such as superconducting and photonic) and slow-clock (such as neutral atom and ion trap) architectures. Our analysis reveals that the first fast-clock CRQCs would enable on-spend attacks on public mempool transactions of some cryptocurrencies. We survey major cryptocurrency vulnerabilities through this lens, identifying systemic risks associated with advanced features in some blockchains such as smart contracts, Proof-of-Stake consensus, and Data Availability Sampling, as well as the enduring concern of abandoned assets. We argue that technical solutions would benefit from accompanying public policy and discuss various frameworks of digital salvage to regulate the recovery or destruction of dormant assets while preventing adversarial seizure. We also discuss implications for other digital assets and tokenization as well as challenges and successful examples of the ongoing transition to Post-Quantum Cryptography (PQC). Finally, we urge all vulnerable cryptocurrency communities to join the ongoing migration to PQC without delay.

Comments: 57 pages, 14 figures

Subjects: Quantum Physics (quant-ph), Cryptography and Security (cs.CR)

Cite as: arXiv 2603.28846 [quant-ph]

(or arXiv:2603.28846v1 [quant-ph] for this version)

<https://doi.org/10.48550/arXiv.2603.28846>

Published today on arxiv.org by combined team: Google Quantum AI, Santa Barbara, CA 93111, United States
2Department of Computer Science, University of California Berkeley, Berkeley, CA 94720, United States
3Ethereum Foundation, Zeughausgasse 7a, 6300 Zug, Switzerland



In System Weakness by Berend Watchus

The Quiet Exit: How a Private Lab May Have Already Broken Blockchain Cryptography — and Why You...

Author: Berend Watchus. Publication for SYSTEM WEAKNESS. April 1, 2026

Apr 1

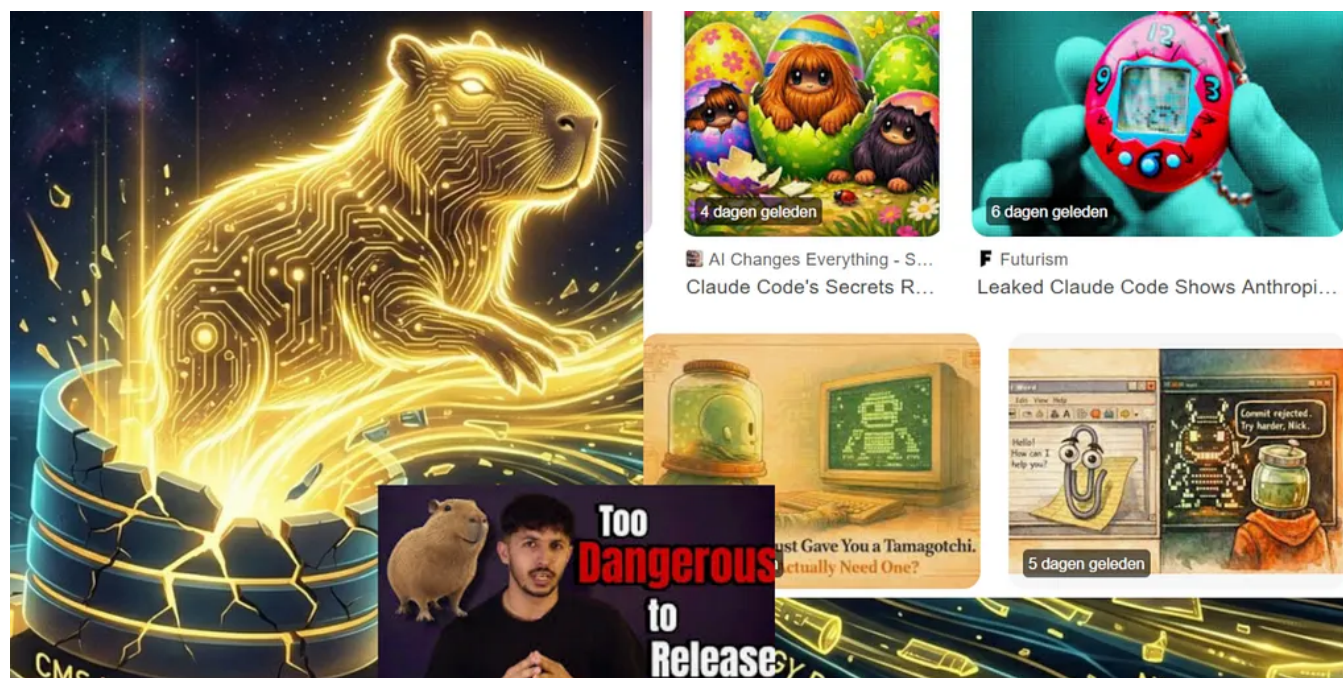


 In OSINT Team by Berend Watchus

“It Wasn’t Me”: RELIABILITY AND ADMISSIBILITY OF AI-GENERATED FORENSIC EVIDENCE IN CRIMINAL TRIALS

Author: Berend Watchus Independent non profit AI & Cyber Sec Researcher [Publication for: OSINT Team, online magazine]

Jan 14  73



 In OSINT Team by Berend Watchus

How Anthropic's Leak Became a Meme Storm Nobody Planned (Part 2)

Author: By Berend F. Watchus. Independent AI & Cyber Security Researcher. Publication for OSINT Team

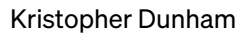
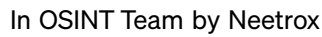
Apr 8  50



See all from Berend Watchus

Recommended from Medium



[illegible]

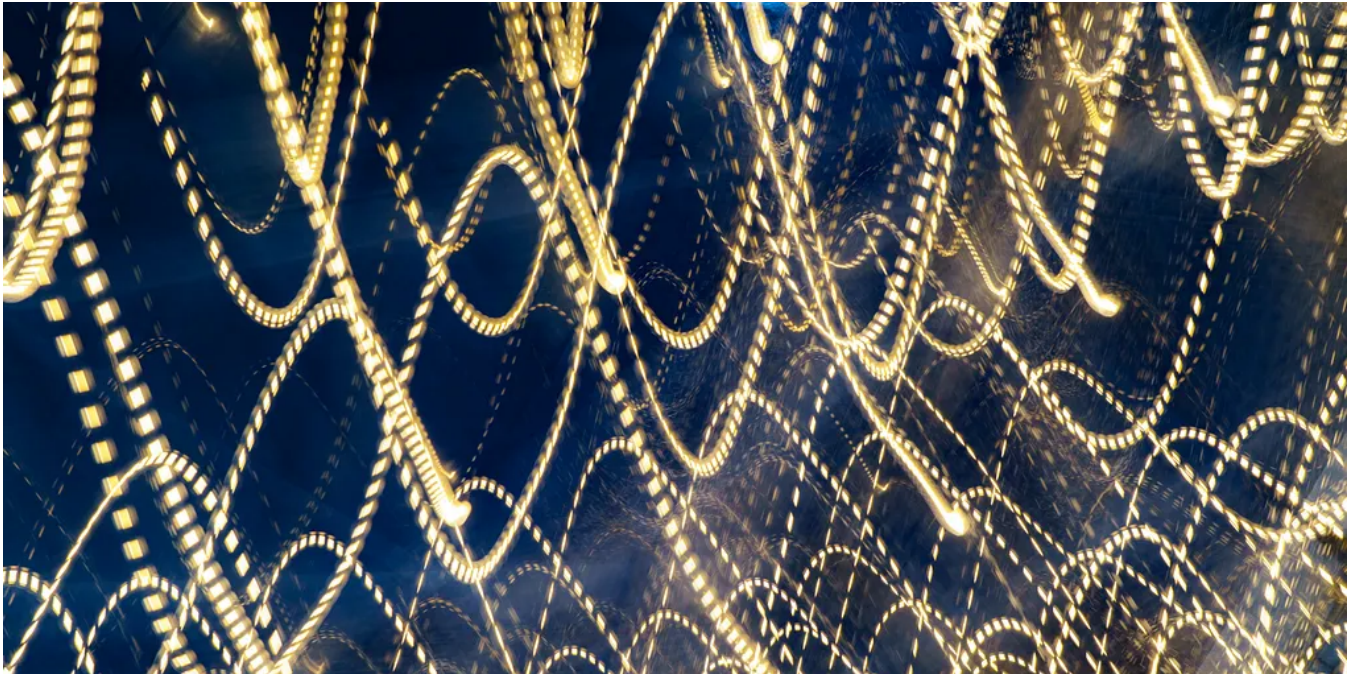
Stop Scrolling Cybersecurity News: Let This AI Send You Only the Threats That Matter

★ Mar 17 🖱️ 156

Stop Scrolling Cybersecurity News: Let This AI Send You Only the Threats That Matter

156





Fatealy

8 Wireshark Patterns That Instantly Signal Something Is Wrong

Once you see them, you can't unsee them



Mar 31



194



1



Linux
Hidden
Directories

 bektiaw

9 Hidden Linux Directories Every Engineer Should Know

They're out of sight, but they quietly control how your system behaves.

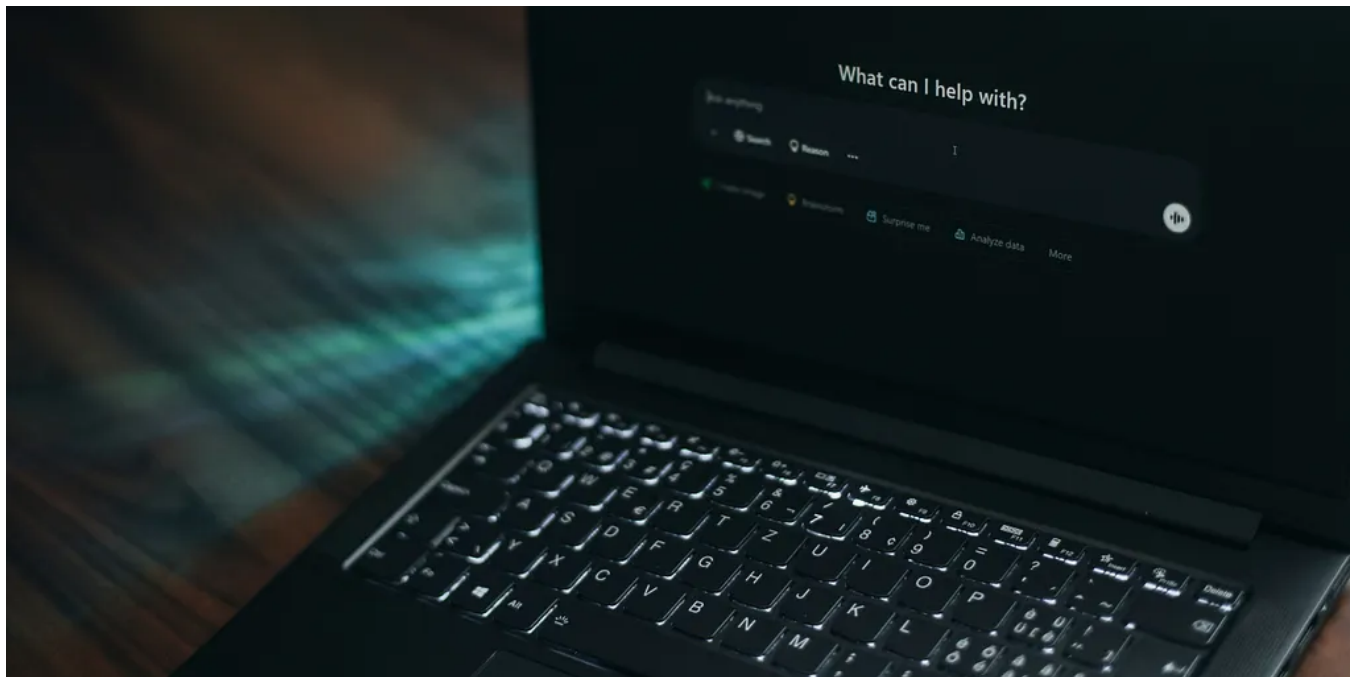
 Siva Sankar

Dark Web Hunting & Monitoring: A Practical Guide for Pentesters & Red Teamers — Part 6

Exploring the Darkweb

★ Apr 8 🖐️ 16 💬 1





In Write A Catalyst by Mubashir

Ethical Hacking: Where to Start Legally

You do not have to do anything to learn about keeping things secure. Here is how you can build skills in a good way. With approval a reason...



Apr 2



92



See more recommendations